데이터마이닝개론 Assignment 8

20191012 김효림

1. 기존에 있었던 데이터 셋 기반 숙제 문제 중 하나를 선정하여서 Neural Network를 적용시키시오.

- 사용한 데이터셋: student_health_2.csv
- 기존에 사용한 기법: KNN
- 새로 적용한 기법: NN
- 코드는 별도의 (.py)형식의 첨부파일로 제출

2. 해당 문제에 썼던 기법과 Neural Net 과의 차이점을 결과값으로 보여주고 시사점을 작성하시오.

2.1. 데이터의 구성 확인 및 독립변수 선정

student_health_2.csv 파일은 size (3569, 25)의 데이터를 가진다. 데이터 활용 목적은 학년을 예측하기 위함이며, 이 종속변수에 영향을 미치는 feature는 아래의 '그림1'과 같다. 그러나 '혈당식전mgdl', '총콜레스테롤mgdl', 'ASTUL', 'ALTUL', '혈색소gdl', '간염검사'의 경우 null 값이 대부분이었으며, '수축기'와 '이완기'의 경우 1학년과 4학년 학생들에 한정하여 데이터가 존재한다. 이에 따라, 아래의 두 가지 상황에서 KNN과 NN을 각각 적용해보고 차이를 비교해보고자 한다.

키	몸무게	혈당식전m	총콜레스E	ASTUL	ALTUL	혈색소gdl	간염검사	수축기	이완기

그림 1 학년에 영향을 미치는 독립변수

2.2 비교 기준

a. 정확률: sklearn 내장함수인 score에 동일한 test_data를 넣어서 도출한 값.

b. 훈련 속도: sklearn 내장함수인 fit에 동일한 train_data를 넣어서 도출한 값.

단, 두 모델 모두 별도의 파라미터를 지정하지 않은 디폴트 값으로 train을 진행한다.

2.3 비교 상황

a. 2개의 feature('키', '몸무게')로 6개의 Category(1~6학년) 예측하기.

b. 4개의 feature('키', '몸무게', '수축기', '이완기')로 2개의 Category(1, 4학년) 예측하기.

2.4 결과 비교 및 시사점

Case 1) 2개의 feature로 6개 카테고리 분류

	KNN	NN
훈련 속도	0.00149	0.35464
정확도(score)	0.45352	0.30459
Test Data Predict 결과	[2 6 5 3 1]	[3 3 6 3 3]

Case 2) 4개의 feature로 2개 카테고리 분류

	KNN	NN
훈련 속도	0.00192	0.35085
정확도(score)	0.96453	0.83333
Test Data Predict 결과	[4 4 1 4 4]	[4 4 1 4 4]

두 개의 케이스 모두에서 KNN 보다 NN의 훈련 속도가 300배 이상 오래 걸렸다. 반면, 정확도의 경우 KNN이 15%p 정도 높음을 확인할 수 있다. 단편적으로 보았을 때는 훈련시간은 적게 걸리나 정확도가 높은 KNN이 더 좋은 모델로 평가될 수 있으나, 각 모델은 각각의 장단점에 기반하여 용도를 달리한다.

KNN의 경우 카테고리별로 구분이 명확한 형식의 데이터를 예측할 경우, 빠른 train 시간과 높은 정활도를 보일 수 있으나, 카테고리의 경계가 모호한 데이터의 경우에는 데이터 예측에 여려움이 있을 수 있다. 조정 가능한 파라미터도 한정적이다.

반면, NN의 경우 hidden layer의 수, drop out 계층, batch size 등의 다양한 하이퍼 파라미터를 조정함으로써, 보다 다양한 상황에서 정교하게 예측을 할 수 있다는 장점을 가진다. 이를 통해 Training 시간을 조정할 수도, 성능을 향상시킬 수 있다.